JOURNAL OF COMPUTER AND SYSTEM SCIENCES 8, 220–237 (1974)

Solution of Eigenvalue Problems in Hilbert Spaces by a Gradient Method*

E. K. BLUM AND G. H. RODRIGUE†

Department of Mathematics, University of Southern California, Los Angeles, California 90007

Received August 13, 1973

A gradient technique is developed for computing a class of nonisolated stationary points, called C-stationary points, for a real functional F defined on a Hilbert space.

It is shown that the least-squares solutions of the operator equation $Ax = b$ are C-stationary points for the functional $(1/2) \| Ax - b \|^2$ when $R(A)$ is closed and that certain eigenvectors of the general eigenproblem $Ax = \lambda Bx$ are C-stationary points for the functional $\frac{1}{2} \| Ax - (\langle Ax, Bx \rangle / \langle Bx, Bx \rangle) Bx \|^2$. Numerical experiments are given to justify the results.

INTRODUCTION

In this paper a gradient technique is developed for computing a class of nonisolated stationary points, called "C-stationary" points, for a real functional F defined on a Hilbert space. In Theorem 1, the method is shown to converge linearly to a C-stationary point provided an appropriate initial vector is chosen. Many of the ideas leading to Theorem 1 were furnished in the paper by Blum [2] where one finds a strong similarity between the quasiregular points defined there and the C-stationary points defined here. In Theorem 2, a weakened monotonicity assumption on the second derivative is shown to be a sufficient condition for a linear variety to be C-stationary. A more complete development of this idea can be found in McCormick [7].

In Section 2, the results of Section 1 are applied to the least-squares problem. It is shown that the existence of a continuous pseudoinverse is sufficient to guarantee that the set of least-squares solutions is C-stationary.

In Section 3, the results of Section 1 are applied to the general eigenproblem $Ax = \lambda Bx$. Here, conditions on the spectrum are presented which force an eigenspace to be C-stationary. Many of the results in Sections 2 and 3, were taken from Rodrigue [11].

* This research was supported by NSF Grant GP 20130.

† Present address: Dept. of Mathematics, Kent State University, Kent, Ohio.

1. C-Stationary Sets

At the outset, we wish to stress that we are concerned in general with stationary points rather than extrema of functionals. A stationary point $\tilde{x}$ of a functional F is a point where some type of derivative $F'(\tilde{x}) = 0$. (We shall take $F'(x)$ to be the Frechet (or strong) derivative.) Of course, a local extremum point at which F has a derivative must be a stationary point, but certainly the converse need not hold. Many existing gradient methods are iterative methods for determining local extrema, rather than stationary points in general. (See [1, 3, and 10].) Thus, ascent or descent methods may not be applicable. Furthermore, we are interested in stationary points which are not isolated. In fact, in Sections 2 and 3 we consider functionals for which the set of stationary points is a linear variety (with possibly one point excluded). We present a gradient method which yields a sequence (x_n) which converges to such nonisolated stationary points when the functional F is question satisfies certain easily motivated conditions. The central idea is that the angle which the gradient vector $\nabla F(x)$ makes with the vector $x - \tilde{x}$, where $\tilde{x}$ is the nearest stationary point to x, should be bounded away from $\pi/2$ by a fixed amount in a certain neighborhood. Note that this differs from the angles considered in [10], where directions other than gradient directions are considered, as in conjugate direction (variable metric) methods. Another point of difference between our method and various other gradient techniques [3], is that we choose the step size s so that x_{n+1} is nearly the closed point on the line $x_n + s\,\nabla F(x_n)$ to the set of stationary points. (This is made precise below.) Hence, our gradient method could be called the "method of closest approach." Since we are interested in infinite-dimensional problems, we present the method in a Hilbert space setting.

Let H be a Hilbert space with real inner-product $\langle x, y\rangle$ and F a real functional defined on H. The strong gradient of F at $x \in H$ is denoted by $\nabla F(x)$. A subset E of H is said to be a *stationary set* of F if for all $x \in E$, $\nabla F(x) = 0$.

We adopt the notation $\bar{x}$ for the unit vector $x/\| x \|$, $x \neq 0 \in H$.

DEFINITION 1. A stationary set E of F is said to be a *C-stationary set for F* if for every $\epsilon > 0$ there exists a neighborhood N, of E, and a constant $a > 0$ such that the following conditions are satisfied:

For $x \in N$,

(1) $\nabla F(x)$ exists as a continuous function of x;

(2) There exists a unique nearest point $\tilde{x} \in E$ for which

$$\| x - \tilde{x} \| = \inf_{w \in E} \| w - x \|;$$

For $x \in N - E$,

(3) $\nabla F(x) \neq 0$;

(4) Let $\tilde{x}$ be the nearest point to x in E and $\Delta x = x - \tilde{x}$.

Define

$$\cos \varphi(x) = \langle \overline{\nabla F(x)}, \overline{\Delta x} \rangle.$$

Then

$$\cos^2 \varphi(x) \geqslant a \tag{4a}$$

and

$$\left| \cos \varphi(x) - \frac{2(F(x) - F(\tilde{x}))}{\| \nabla F(x) \| \, \| \Delta x \|} \right| < \epsilon. \tag{4b}$$

N is called a *C-stationary neighborhood* of E.

We adopt the notation $\| x - E \| = \inf_{w \in E} \| x - w \|$.

Among the principal methods for finding stationary points are those which involve moving in the direction of the gradient of F. Furthermore, in certain problems where it is required to find the zeros of F, the problem can be reformulated so that the zero is also a stationary point. F may be the norm-squared of the gradient of another functional. Two important examples of this kind of reformulation are given.

The various gradient techniques differ in their choice of step-sizes, for example, the steepest descent method and the gradient methods given in [2, 3, and 10]. In the gradient method presented here, the choice of step-size is a linear estimate of the distance along the gradient from a point x to the point nearest to the closest stationary point. To be more precise, let $\tilde{x}$ be the nearest stationary point to a vector x. The value of s which minimizes the quantity

$$\| x - s \nabla F(x) - \tilde{x} \|^2$$

is given by

$$s = \langle \nabla F(x), x - \tilde{x} \rangle / \| \nabla F(x) \|^2.$$

Expanding F in a Taylor series about $\tilde{x}$, we obtain (with $h = x - \tilde{x}$),

$$F(x) = F(\tilde{x}) + (1/2) F''(\tilde{x}) h^2 + o(h^2).$$

In a similar fashion, we expand $F'(x)$ in a Taylor series about $\tilde{x}$ to obtain

$$F'(x)h = \langle \nabla F(x), h \rangle = F''(\tilde{x}) h^2 + o(h^2).$$

Combining this with the first expansion we obtain

$$F(x) = F(\tilde{x}) + (1/2) \langle \nabla F(x), h \rangle + o(h^2).$$

Letting $h = x - \tilde{x}$ and ignoring terms $o(h^2)$, we obtain

$$\langle \nabla F(x), x - \tilde{x} \rangle / \| \nabla F(x) \|^2 \simeq 2[F(x) - F(\tilde{x})] / \| \nabla F(x) \|^2$$

The term on the right of the above relation is the step-size which we will using. Note that it becomes a computable step-size when the stationary points of F are also

zero points, i.e., $F(\tilde{x}) = 0$ or when $F(\tilde{x})$ is known. In order to achieve generality, we will not assume that stationary points are zero points.

THEOREM 1. *Let E be a C-stationary set for F. For an arbitrary C-stationary neighborhood N, of E, let $\tilde{x} \in E$ be the unique nearest point to $x \in N$ and define*

$$h(x) = \begin{cases} \dfrac{-2(F(x) - F(\tilde{x}))}{\| \nabla F(x) \|^2} \nabla F(x) & , \quad \text{if} \quad \nabla F(x) \neq 0; \\ 0, & \text{if} \quad \nabla F(x) = 0. \end{cases}$$

Then there exists a neighborhood M, of E, and a positive constant $k < 1$ such that for any $x_0 \in M$, the sequence defined by

$$x_{n+1} = x_n + h(x_n), \qquad n = 0, 1, 2, \ldots, \tag{5}$$

converges to a point in the closure of E and satisfies the relation

$$\| x_n - E \| \leqslant k^n \| x_0 - E \|. \tag{6}$$

Furthermore, for arbitrary $\eta > 0$, M can be chosen so that

$$| k^2 - (1 - \inf_{x \in M - E} \cos^2 \varphi(x)) | < \eta. \tag{7}$$

Proof. Let N be a C-stationary neighborhood for F. From Definition 1 there exists an $a > 0$ such that

$$\inf_{x \in N - E} \cos^2 \varphi(x) \geqslant a.$$

Let $\eta > 0$ be given. Choose $\epsilon > 0$ such that $\epsilon < 1$ and $\epsilon^2 < \min(\eta, a)$ and let N_ϵ be a corresponding C-stationary neighborhood of E such that $N_\epsilon \subset N$. Let $k^2 = 1 - a + \epsilon^2$. It follows that $0 < k < 1$. Let $x \in N_\epsilon$ be fixed and $\tilde{x} \in E$ the nearest point to x. It then follows from (4) that for $\Delta x = x - \tilde{x}$,

$$\begin{aligned} \| x + h(x) - E \|^2 &\leqslant \| x + h(x) - \tilde{x} \|^2 \\ &= \| \Delta x \|^2 + 2\langle h(x), \Delta x \rangle + \| h(x) \|^2 \\ &\leqslant \| \Delta x \|^2 + \| \Delta x \|^2 (-\cos^2 \varphi(x) + \epsilon^2) \\ &\leqslant \| \Delta x \|^2 (1 - a + \epsilon^2). \end{aligned}$$

Hence

$$\| x + h(x) - E \|^2 \leqslant k^2 \| x - E \|^2. \tag{8}$$

Since N_ϵ is an open set containing E, then for each $e \in E$, there exists an $r_e > 0$ such that the open ball $B(e; r_e) = \{x \in H: \| x - e \| < r_e\}$ is contained in N_ϵ. Let

$$M = \bigcup_{e \in E} B(e; r_e(1 - k)/4).$$

Clearly $M \subset N_\epsilon$. Furthermore, if $x \in B(e; r_e(1-k)/4)$, then $B(x; r_e/2) \subset N_\epsilon$. Let $x_0 \in M$. Then $x_0 \in B(z; r_z(1-k)/4)$ for some $z \in E$. Let $\tilde{x}_0$ be the nearest point in E to x_0 so that $\| x_0 - \tilde{x}_0 \| \leqslant \| x_0 - z \|$. We now wish to establish by induction that the sequence (x_n) given by (5) and the corresponding nearest points $\{\tilde{x}_n\} \subset E$ satisfy the properties:

(i) $x_{n+1} \in B(x_0\,;\, \delta_0)$, $\delta_0 = r_z/2$;
(ii) $\| x_{n+1} - \tilde{x}_{n+1} \| \leqslant k^{n+1} \| x_0 - \tilde{x}_0 \|$;
(iii) $\| x_{n+1} - x_n \| \leqslant 2 \| x_n - \tilde{x}_n \|$;
(iv) $\| \tilde{x}_{n+1} - \tilde{x}_n \| \leqslant (k+3) \| x_n - \tilde{x}_n \|$.

It will then follow from (i) that all $x_n \in B(z; r_z) \subset N_\epsilon$. Applying induction to (8) we then obtain the relation (6). Properties (ii), (iii), and (iv) will be used to prove that x_n actually converges to a point. For $n = 1$, since $x_0 \in N_\epsilon$, it follows from (8) that

$$\| x_1 - \tilde{x}_1 \| \leqslant k \| x_0 - \tilde{x}_0 \|.$$

Further,

$$\begin{aligned} \| x_1 - x_0 \| &= 2 \, | F(x_0) - F(\tilde{x}_0) | / \| \nabla F(x_0) \| \\ &= \| x_0 - \tilde{x}_0 \| \, | \cos \varphi(x_0) + 2[F(x_0) - F(\tilde{x}_0)] / \| \nabla F(x_0) \| \, \| x_0 - \tilde{x}_0 \| \\ &\quad - \cos \varphi(x_0) | \\ &\leqslant \| x_0 - \tilde{x}_0 \| \, (| \cos \varphi(x_0) | + \epsilon) \\ &\leqslant \| x_0 - \tilde{x}_0 \| \, (1 + \epsilon) \\ &\leqslant 2 \| x_0 - \tilde{x}_0 \|. \end{aligned}$$

Hence,

$$\| x_1 - x_0 \| \leqslant r_z(1-k)/2 < \delta_0,$$

i.e., $x_1 \in B(x_0\,;\, \delta_0) \subset N_\epsilon$. Further,

$$\begin{aligned} \| \tilde{x}_1 - \tilde{x}_0 \| &\leqslant \| \tilde{x}_1 - x_1 \| + \| x_1 - x_0 \| + \| x_0 - \tilde{x}_0 \| \\ &\leqslant k \| x_0 - \tilde{x}_0 \| + 2 \| x_0 - \tilde{x}_0 \| + \| x_0 - \tilde{x}_0 \| \\ &= (k+3) \| x_0 - \tilde{x}_0 \|. \end{aligned}$$

This establishes properties (i)–(iv) for $n = 0$. Suppose they are true for $n = i$. Then since $x_i \in N_\epsilon$, we have again by (8)

$$\begin{aligned} \| x_{i+1} - \tilde{x}_{i+1} \| &\leqslant k \| x_i - \tilde{x}_i \| \\ &\leqslant k^{i+1} \| x_0 - \tilde{x}_0 \|. \end{aligned}$$

Applying the same argument as before we obtain

$$\| x_{i+1} - x_i \| \leqslant 2 \| x_i - \tilde{x}_i \|$$

and

$$\| \tilde{x}_{i+1} - \tilde{x}_i \| \leqslant (k+3) \| x_i - \tilde{x}_i \|,$$

thus establishing (iii) and (iv). Using the induction hypothesis again,

$$\begin{aligned}
\| x_{i+1} - x_0 \| &= \Big\| \sum_{j=0}^{i} (x_{j+1} - x_j) \Big\| \\
&\leqslant \sum_{j=0}^{i} 2 \| x_j - \tilde{x}_j \| \\
&\leqslant 2 \sum_{j=0}^{i} k^j \| x_0 - \tilde{x}_0 \| \\
&\leqslant \frac{2}{1-k} \| x_0 - \tilde{x}_0 \|. \\
&\leqslant \frac{2}{(1-k)} \frac{r_z(1-k)}{4} \\
&= \delta_0
\end{aligned}$$

Hence $x_{i+1} \in B(x_0 ; \delta_0)$ establishing (i). Property (ii) follows from (8).

We now establish that the sequences (x_n) and $(\tilde{x}_n)$ are Cauchy. Let $\omega > 0$ be given. Let I be an integer such that

$$\frac{k^I(k+3)}{2} \delta_0 < \omega.$$

Then for $i \geqslant j \geqslant I$,

$$\begin{aligned}
\| x_i - x_j \| &= \Big\| \sum_{m=j}^{i-1} (x_{m+1} - x_m) \Big\| \\
&\leqslant \sum_{m=j}^{i-1} \| x_{m+1} - x_m \| \\
&\leqslant 2 \sum_{m=j}^{i-1} \| x_m - \tilde{x}_m \| \\
&\leqslant 2 \| x_0 - \tilde{x}_0 \| \sum_{m=j}^{i-1} k^m \\
&= 2 \| x_0 - \tilde{x}_0 \| k^j \sum_{m=0}^{i-j-1} k^m \\
&\leqslant \delta_0 k^j \\
&< \omega.
\end{aligned}$$

For the same integers i, j,

$$\begin{aligned}
\| \tilde{x}_i - \tilde{x}_j \| &= \left\| \sum_{m=j}^{i-1} (\tilde{x}_{m+1} - \tilde{x}_m) \right\| \\
&\leqslant \sum_{m=j}^{i-1} \| \tilde{x}_{m+1} - \tilde{x}_m \| \\
&\leqslant (k+3) \sum_{m=j}^{i-1} \| x_m - \tilde{x}_m \| \\
&\leqslant (k+3) \| x_0 - \tilde{x}_0 \| \sum_{m=j}^{i-1} k^m \\
&= (k+3) \| x_0 - \tilde{x}_0 \| k^j \sum_{m=0}^{i-j-1} k^m \\
&\leqslant \frac{k+3}{1-k} \| x_0 - \tilde{x}_0 \| k^I \\
&< \omega.
\end{aligned}$$

Since (x_n) and $(\tilde{x}_n)$ are Cauchy sequences, they both converge to limit points x' and $\tilde{x}'$, respectively. Property (ii) asserts that

$$\| x_n - \tilde{x}_n \| \leqslant k^n \| x_0 - \tilde{x}_0 \|$$

for all n. Hence,

$$\lim_{n \to \infty} \| x_n - \tilde{x}_n \| = \| x' - \tilde{x}' \| = 0,$$

i.e., $x' = \tilde{x}'$. Note that $\tilde{x}'$ is in the closure of E.

Remark. If ∇F is continuous on the boundary of the C-stationary set E, then under the same hypothesis of Theorem 1 the sequence of iterates generated by (5) converges to a stationary point. It is of course not always true that ∇F is continuous on the boundary of a C-stationary set as will be observed in Section 3 dealing with eigenvalue problems where ∇F is not continuous at the origin.

We now examine special nonisolated stationary points, namely, those that form a subset of a linear variety. Examples of these types of problems are the subject of Sections 2 and 3.

A subset $M \subset H$ is called a closed linear variety if $M = x_0 + M_0$ where x_0 is a

fixed vector and M_0 a closed linear subspace of H. Since a closed linear variety M is convex, if follows that for all $x \in H$ there exists a unique vector $y \in M$ such that

$$\| x - M \| = \| x - y \|, \tag{9}$$

$$\langle x - y, y - z \rangle = 0, z \in M. \tag{10}$$

THEOREM 2. *Let E be a stationary set for F. In addition, suppose $E \subset M$ where $M = x_0 + M_0$ is a closed linear variety. Let $M_0^{\perp}$ be the orthogonal complement of M_0. Suppose there exists a neighborhood N of E such that*:

(i) *for each $x \in N$ there exists $y \in E$ necessarily unique, such that $\| x - M \| = \| x - y \|$;*

(ii) *for all $x \in N$, the strong second derivative $F''(x)$ exists and is continuous in x;*

(iii) *there exists a constant $c > 0$ such that for every $x \in E$ and every nonzero $h \in M_0^{\perp}$, $\langle F''(x)\bar{h}, \bar{h} \rangle \geqslant c$;*

(iv) *There exists a constant $K > 0$ such that for every $x \in E$ and every $h \in H$, $\| F''(x)h \| \leqslant K \| h \|$;*

(v) *for $x \in N - E$, $\nabla F(x) \neq 0$.*

Then E is a C-stationary set for F.

Proof. To establish the theorem, we construct, for given $\epsilon > 0$, a neighborhood $N(\epsilon) \subset N$ of E such that (4) of Definition 1 is true. The remaining part of the definition will then follow immediately from (i), (ii), and (v). We proceed by first fixing $e \in E$. For $\epsilon_0 > 0$, the continuity of F'' on N implies the existence of a ball $B(e; \delta) \subset N$ such that for any $x, y \in B(e; \delta)$, $\| F''(x) - F''(y) \| < \epsilon_0$. Let $x \in B(e; \delta)$ be such that $x \notin E$. Also, let $\tilde{x} \in E$ be such that $x - \tilde{x} = h$ and $\| h \| = \| x - M \|$. Note that $\tilde{x} \in B(e; \delta)$. Also, it follows from (10) that $h \in M_0^{\perp}$. Using the continuity of F'' on $B(e; \delta)$ and the Taylor expansion of F, we obtain for δ sufficiently small,

$$\begin{aligned} F(x) &= F(\tilde{x}) + 1/2 \langle F''(\tilde{x} + \theta h)h, h \rangle \\ &= F(\tilde{x}) + 1/2 \langle F''(\tilde{x})h, h \rangle + 1/2 \langle \epsilon_1 h, h \rangle, \end{aligned} \tag{11}$$

and

$$\begin{aligned} \nabla F(x) &= F''(\tilde{x})h + \epsilon_2 , \\ &= F''(\tilde{x})h + \epsilon_3 \| h \|, \end{aligned} \tag{12}$$

where $0 < \theta < 1$, and $\max(\| \epsilon_1 \|, \| \epsilon_3 \|) < \epsilon_0$. Using (12) and (iv),

$$\| \nabla F(x) \| \leqslant K \| h \| + \| \epsilon_2 \| \leqslant (K + \| \epsilon_3 \|) \| h \|. \tag{13}$$

It thus follows from (13) and (iii) that

$$\begin{aligned}\cos\varphi(x) &= \frac{\langle \nabla F(x), \bar{h}\rangle}{\| \nabla F(x)\|} \\ &= \frac{(\langle F''(\tilde{x})\bar{h}, \bar{h}\rangle + \langle \epsilon_2 , \bar{h}\rangle)\| h \|}{\| \nabla F(x)\|} \\ &\geqslant \frac{\langle F''(\tilde{x})\bar{h}, \bar{h}\rangle}{K + \epsilon_0} + \frac{\langle \epsilon_2 , \bar{h}\rangle}{K + \epsilon_0} \\ &\geqslant \frac{c}{K + \epsilon_0} + \frac{\langle \epsilon_2 , \bar{h}\rangle}{K + \epsilon_0}. \end{aligned} \tag{14}$$

Hence, by taking ϵ_0 sufficiently small, we obtain a neighborhood N_1 of E and a constant $a > 0$ such that $\cos\varphi(x) \geqslant a$ for $x \in N_1 - E$. Combining (11) and (12),

$$\begin{aligned}\frac{2[F(x) - F(\tilde{x})] - \langle \nabla F(x), h\rangle}{\| \nabla F(x)\| \| h \|} &= \frac{\langle \epsilon_1 h, h\rangle - \langle \epsilon_2 , h\rangle}{\| \nabla F(x)\| \| h \|} \\ &\leqslant \frac{2\epsilon_0}{c + \langle \epsilon_3 , \bar{h}\rangle}\end{aligned}$$

since $\| \nabla F(x)\| \| h \| \geqslant \langle \nabla F(x), h\rangle \geqslant \| h \|^2 (c + \langle \epsilon_3 , \bar{h}\rangle)$. Hence, if we take $\epsilon_0 > 0$ so that $| 2\epsilon_0/(c + \langle \epsilon_3 , \bar{h}\rangle)| < \epsilon$, we obtain a neighborhood $N(\epsilon) \subset N_1$ such that $E \subset N(\epsilon)$ and (4a), (4b) holds. ∎

Remark 1. It follows from (14) that for arbitrary $\epsilon > 0$ the corresponding C-stationary neighborhood $N(\epsilon)$ of E can be chosen so that for $x \in N(\epsilon) - E$

$$\left| \cos^2\varphi(x) - \frac{\langle F''(\tilde{x})\bar{h}, \bar{h}\rangle^2}{\| F''(\tilde{x})\bar{h} \|^2} \right| < \epsilon.$$

Remark 2. If we drop the condition that E be part of a linear variety, then Theorem 2 still holds if, for all nonzero $h \in H$ and $x \in E$, we demand that $\langle F''(x)\bar{h}, \bar{h}\rangle \geqslant c > 0$. (This makes F a convex functional on E when E is a convex set.)

2. The Linear Least-squares Problem

We consider a bounded linear operator A mapping a Hilbert space H into a Hilbert space H'. Let the range of A, $R(A)$, be closed. It is known (e.g., see Blum [1, Section 7.4]) that a least-squares solution of the equation

$$Ax = b, \quad b \in H' \tag{15}$$

exists and is a solution of the normal equation $A^*Ax = A^*b$. (A^* is the adjoint of A.) Also, if x and x_0 are two solutions of the normal equation, then $x - x_0 \in N$, where N is the null space of A. Thus, any solution of the normal equation minimizes the functional

$$F(x) = (1/2) \| Ax - b \|^2. \tag{16}$$

Let E be the set of least-squares solutions of (15). Then $E = x_0 + N$ where x_0 is any element of E. To apply the previous results to this problem, we observe that

$$\nabla F(x) = A^*(Ax - b)$$

and

$$F''(x) = A^*A.$$

Hence, $\nabla F(x) = 0$ if and only if $x \in E$. Since $R(A)$ is closed it is known that the restriction of A to $N^\perp$ is a bijection of $N^\perp$ onto $R(A)$. Hence this operator has a bounded inverse, i.e., the pseudoinverse of A. Equivalently, there exists a constant $c > 0$ such that for every non-zero $h \in N^\perp$,

$$\langle F''(x)h, h \rangle = \| Ah \|^2 \geqslant c.$$

Since A is a bounded linear operator by assumption, it follows from Theorem 2 that E is a C-stationary set for F. In turn, Theorem 1 implies the existence of a neighborhood N of E for which iteration (5) converges at a linear rate to E. It can be shown (cf. [8]) that the neighborhood N can be taken to be the entire Hilbert space H, i.e., global convergence occurs as in other gradient methods (cf. [1, Sect. 12.3; 3; 6; 9]. It might be worth pointing out again that in contrast to the method of steepest descent where the step-size at a point x is chosen to minimize the functional F in the direction $\nabla F(x)$, the step-size used for iteration (5) is in this (quadratic) case exactly the value of s which minimizes the quantity $\| x - s \nabla F(x) - \tilde{x} \|^2$ where, as before, $\tilde{x}$ is the least-squares solution nearest to x.

Unfortunately, (5) apparently requires that we know solutions $\tilde{x}_n \in E$ nearest to each x_n. Actually, we only need to know $F(\tilde{x}_n)$, which in this case is the unique value $\mu = \min_x F(x)$. Although μ will usually not be known, an approximation may suffice as the proof of Theorem 2 demonstrates. One important case requires no approximation, namely, the case where we know that $b \in R(A)$ so that $\mu = 0$. An alternative procedure would be to apply the preceding discussion to the functional $F(x) = (1/2) \| A^*(Ax - b) \|^2$ where here $F(x) = 0$ if and only if x is a least-squares solution regardless of whether or not b is an element of $R(A)$. However, until further studies are completed, we do not suggest the use of this method in actual computation.

3. Eigenvalue Problem

Consider the eigenvalue problem

$$Ax = \lambda Bx \tag{17}$$

where A and B are arbitrary bounded linear operators mapping one Hilbert space H into another Hilbert space H'. We assume $Bx \neq 0$ whenever x is a nonzero eigenvector of (17). Consider the following functional

$$F(x) = (1/2) \| Ax - (\langle Ax, Bx\rangle / \langle Bx, Bx\rangle) Bx \|^2, \tag{18}$$

defined for $x \in H$ such that $Bx \neq 0$. Note that the eigenvectors of (17) are precisely the local minima of F. For nonzero x, we have for the differential of F,

$$dF(x; h) = \langle (Ax - R(x) Bx)'h, Ax - R(x) Bx\rangle$$

where $R(x) = \langle Ax, Bx\rangle / \langle Bx, Bx\rangle$ and

$$(Ax - R(x) Bx)'h = (A - R(x)B)h - dR(x; h) Bx.$$

Since $dR(x, h)$ is a scalar and $\langle Bx, Ax - R(x) Bx\rangle = 0$,

$$dF(x; h) = \langle (A - R(x)B)h, (A - R(x)B)x\rangle.$$

Hence,

$$\nabla F(x) = (A - R(x)B)^* (A - R(x)B)x. \tag{19}$$

Using (19), we have $\langle \nabla F(x), x\rangle = 2F(x)$ so that $\nabla F(x) = 0$ if and only if $F(x) = 0$, that is, the only stationary points of F are eigenvectors of (17) and conversely. Note, of course, that eigenvalues and eigenvectors need not exist. We now show that (iii) and (iv) of Theorem 2 hold.

For $x, h \in H$ and s a real scalar, let $z(s) = \nabla F(x + sh)$. It then follows that $F''(x) \cdot h = z'(0)$.

Let $T(x) = (A - R(x)B)^* (A - R(x)B)$. It follows from (19) that (with $'$ denoting differentiation with respect to s)

$$z'(s) = T'(x + sh) \cdot (x + sh) + T(x + sh) \cdot (x + sh)'$$

so that

$$F''(x) \cdot h = z'(0) = dT(x; h) \cdot x + T(x) \cdot h \tag{21a}$$

Now,

$$\begin{aligned} T'(x + sh) &= [(A - R(x + sh)B)^*]' (A - R(x + sh)B) \\ &\quad + (A - R(x + sh)B)^* (A - R(x + sh)B)' \\ &= -R'(x + sh) B^*(A - R(x + sh)B) \\ &\quad - (A - R(x + sh)B)^* R'(x + sh)B \end{aligned}$$

so that, setting $s = 0$, we obtain for the differential,

$$dT(x; h) = -dR(x; h)[B^*(A - R(x)B) + (A - R(x)B)^* B].$$

Upon expanding,

$$R(x + sh) = \frac{\langle Ax, Bx\rangle + s(\langle Ax, Bh\rangle + \langle Ah, Bx\rangle) + s^2\langle Ah, Bh\rangle}{\| Bx\|^2 + 2s\langle Bx, Bh\rangle + s^2 \| Bh\|^2}.$$

Hence,

$$R'(x)h = \frac{\| Bx\|^2 (\langle Ax, Bh\rangle + \langle Ah, Bx\rangle) - 2\langle Ax, Bx\rangle\langle Bx, Bh\rangle}{\| Bx\|^4}$$

$$= \| Bx\|^{-2} \langle (A^*B + B^*A)x, h\rangle - 2 \| Bx\|^{-4} \langle Ax, Bx\rangle\langle B^*Bx, h\rangle.$$

Let E_λ be the eigenspace corresponding to an eigenvalue λ of (17). Then for nonzero x in E_λ, we have $R(x) = \lambda$ so that for such an eigenvector,

$$T(x) = (A - \lambda B)^* (A - \lambda B), \tag{21b}$$

$$dT(x; h) \cdot x = -dR(x; h)(A - \lambda B)^* Bx, \tag{21c}$$

and

$$dR(x; h) = \| Bx\|^{-2} [\langle (A + \lambda B)^* Bx, h\rangle - 2\lambda\langle B^*Bx, h\rangle]$$

$$= \| Bx\|^{-2} \langle (A - \lambda B)^* Bx, h\rangle. \tag{21d}$$

It thus follows from (21a, b, c, d) that for nonzero x in E_λ,

$$F''(x)h = - \| Bx\|^{-2} \langle (A - \lambda B)^* Bx, h\rangle (A - \lambda B)^* Bx + (A - \lambda B)^* (A - \lambda B)h \tag{22}$$

and

$$\langle F''(x)h, h\rangle = \|(A - \lambda B)h\|^2 - \langle (A - \lambda B)h, \overline{Bx}\rangle^2. \tag{23}$$

It follows from (22) that there exists a $K > 0$ such that for all $x \in E_\lambda - \{0\}$, $\| F''(x)h\| \leqslant K \| h\|$. This establishes (iv). To establish (iii) we require further assumptions.

ASSUMPTION 1. *Henceforth, we assume that λ is an isolated eigenvalue of* (17).

Let $S = \{x \in H : \| Bx\| = 1\}$. We claim that since λ is an isolated eigenvalue, there exists $0 < d < 1/\| B\|$ such that $Q(d) = \bigcup_{e \in S \cap E_\lambda} B(e; d)$ has an empty intersection with E_μ, $\mu \neq \lambda$. If not, there exists a sequence of eigenvectors $x_n \in S$ corresponding to eigenvalues $\mu_n \neq \lambda$ and a sequence of eigenvectors $e_n \in S \cap E_\lambda$ such that $\| x_n - e_n\| \to 0$ as $n \to \infty$. This implies $\| A(x_n - e_n)\|^2 = (\lambda - \mu_n)^2 + \epsilon_n$ where $\epsilon_n \to 0$. Hence, $\mu_n \to \lambda$, contradicting Assumption 1.

Define $N(d)$ to be the cone $\{\rho Q(d): \rho > 0\}$. It is easily seen from the above discussion that (i), (ii), and (v) of Theorem 2 hold for the neighborhood $N(d)$ with respect to the stationary set $E_\lambda - \{0\}$ (we take $M = E_\lambda$). Hence, in order to establish when $E_\lambda - \{0\}$ will be a C-stationary set for F, it will suffice to establish condition (iii), that

is, for arbitrary $x \in E_\lambda - \{0\}$ and nonzero $h \in E_\lambda^\perp$, $\langle F''(x)h, h\rangle$ will be bounded away from zero.

ASSUMPTION 2. *To obtain* (iii), *we now assume that* B^{-1} *exists and is continuous.*

Remark. Let $T_\lambda = A - \lambda B$. Observe that

$$B^{-1}(T_\lambda E_\lambda^\perp \cap BE_\lambda) = (B^{-1}A - \lambda I)\, E_\lambda^\perp \cap E_\lambda .$$

Also note that

$$(B^{-1}A - \lambda I)\, E_\lambda^\perp \subset \text{Range}(B^{-1}A - \lambda I) \qquad \text{and} \qquad E_\lambda = \text{Kernel}(B^{-1}A - \lambda I).$$

It is known (cf. [12, p. 306]) that if λ is a simple pole of the resolvent of $B^{-1}A$, then

$$H = \text{Kernel}(B^{-1}A - \lambda I) \oplus \text{Range}(B^{-1}A - \lambda I).$$

(As usual, $\oplus$ denotes the direct sum.) This implies that $T_\lambda E_\lambda^\perp \cap BE_\lambda = \{0\}$. Furthermore, for $x \in H'$, there exists y such that $By = x$. Let $y = u + v$, where $u \in E_\lambda$ and $v \in \text{Range}(B^{-1}A - \lambda I)$. Then $x = Bu + Bv$. Since $Bv \in \text{Range}(T_\lambda) = T_\lambda E_\lambda^\perp$ we see that $H' = BE_\lambda \oplus T_\lambda E_\lambda^\perp$.

THEOREM 3. *If* λ *is a simple pole of the resolvent of* $B^{-1}A$, *then there exists a constant* $K > 0$ *such that* $\langle F''(x)h, h\rangle \geqslant K^2 \| h \|^2$, $x \in E_\lambda - \{0\}$, $h \in E_\lambda^\perp$, *i.e.,* $E_\lambda - \{0\}$ *is a C-stationary set.*

Proof. Since B^{-1} exists and is continuous, BE_λ is a closed linear subspace. Let $M = BE_\lambda$ and $M^\perp = (BE_\lambda)^\perp$ and let their corresponding orthogonal projections be P_M, $P_M{}^\perp$.

We claim that $P_M{}^\perp$ maps $T_\lambda E_\lambda^\perp$ bijectively onto $M^\perp$. It follows from the preceding remark that $T_\lambda E_\lambda^\perp \cap BE_\lambda = \{0\}$. Hence, if $P_M{}^\perp x = 0$ for some $x \in T_\lambda E_\lambda^\perp$, then $x = 0$, i.e., $P_M{}^\perp$ is injective on $T_\lambda E_\lambda^\perp$. Let $x \in M^\perp$. Then it follows from the above remark that $x = p + q$ where $p \in T_\lambda E_\lambda^\perp$ and $q \in M$. Then,

$$x = P_M{}^\perp x = P_M{}^\perp p + P_M{}^\perp q = P_M{}^\perp p.$$

It follows that $P_M{}^\perp T_\lambda$ is a bijective map from $E_\lambda^\perp$ onto $M^\perp$. Hence by the theorem of Banach, the inverse of $P_M{}^\perp T_\lambda$ exists and is continuous, or equivalently, there exists a constant $K > 0$ such that $\| P_M{}^\perp T_\lambda h \| \geqslant K \| h \|$, $h \in E_\lambda^\perp$. Using (23),

$$\langle F''(x)h, h\rangle = \| T_\lambda h - \langle T_\lambda h, \overline{Bx}\rangle \overline{Bx} \|^2. \tag{24}$$

Hence,

$$\begin{aligned}\langle F''(x)h, h\rangle &\geqslant \min_{x \in E_\lambda} \| T_\lambda h - \langle T_\lambda h, \overline{Bx}\rangle \overline{Bx} \|^2 \\ &= \|(I - P_M)\, T_\lambda h \|^2 \\ &= \| P_M{}^\perp T_\lambda h \|^2 \\ &\geqslant K^2 \| h \|^2. \quad \blacksquare\end{aligned}$$

In the case when dimension (H) = dimension $(H') < \infty$, it follows (cf. [12, p. 314]) that λ is a simple pole of the resolvent of $B^{-1}A$ if and only if all of the elementary divisors of $B^{-1}A$ corresponding to λ are linear. We then have the following result.

COROLLARY 1. *Suppose dimension* (H) = *dimension* $(H') < \infty$. *If all of the elementary divisors of* $B^{-1}A$ *corresponding to* λ *are linear, then* $E_\lambda - \{0\}$ *is a C-stationary set.*

We remark that the requirement of linear elementary divisors to attain local linear convergence to an eigenspace is not surprising. From a computational standpoint, linear elementary divisors of an eigenvalue λ is necessary for λ to be well-conditioned (cf. [13, p. 154]). Hence, if the term well-conditioned could be properly defined, an equivalent statement of Corollary 1 might be: If λ is a well-conditioned eigenvalue, then $E_\lambda - \{0\}$ is a C-stationary set.

For the matrix case, with A and B arbitrary, the gradient method presented here should be compared with that given recently in [16]. It seems likely that the gradient method may have the advantage in the special case where A and B are band or sparse matrices, since our method uses only A and B and their adjoints, as formulas (18) and (19) show. In the next two sections, we given some computational results which verify the theory.

4. COMPUTATIONAL RESULTS: LEAST-SQUARES PROBLEM

The gradient method described in Theorem 1 as applied to the least-squares problem was programmed in WATFIV on an IBM 360/55 in double-precision. The function F in formula (16) was used. The computing time is noted, however it should be borne in mind that no programming optimization was attempted. It has been proved that convergence is obtained with any initial vector [8].

$$A = \begin{bmatrix} 2 & 1 & 3 & 4 \\ 1 & -3 & 1 & 5 \\ 3 & 1 & 6 & -2 \\ 4 & 5 & -2 & -1 \end{bmatrix}$$

$$b^* = [10 \quad 4 \quad 8 \quad 6]$$

$$x_0^* = [1 \quad 2 \quad 3 \quad 4] \qquad \text{(initial vector)}$$

no. of iterations	true solution	computed solution
177	1	0.999998
	1	1.000002
	1	1.000000
	1	1.000000

CPU time = 1.32 sec.

5. Computational Results: Eigenvalue Problem

The method described in Theorem 1 as applied to the eigenvalue problem was programmed in Fortran IV on an IBM 360/55 in double precision. We choose an arbitrary initial vector x_0, compute $F(x_0)$ using formula (18) and $\nabla F(x_0)$ by (19). Then $h(x_0)$ is given by the formula in Theorem 1 and the next iterate x_1 is given by formula (5). Symmetric matrices A and B are used. In addition, B is positive-definite so that the eigenvectors $\{v_i\}_{i=1}^{k}$ corresponding to distinct eigenvalues satisfy the B-orthogonality property, $\langle Bv_i, v_j\rangle = 0$, $i \neq j$. This suggests the usage of a Gram–Schmidt orthogonalization process to obtain the remaining eigenvectors once the first r-eigenvectors have been computed. To be more precise, suppose r B-orthogonal eigenvectors $\{v_i\}_{i=1}^{r}$ have been computed. The algorithm to compute the $(r+1)$st eigenvector is as follows:

$$y_{n+1} = x_n - \sum_{i=1}^{r} \frac{\langle x_n, Bv_i\rangle}{\langle Bv_i, v_i\rangle} v_i,$$

$$x_{n+1} = x_n + h(y_{n+1}),$$

where h is given by Theorem 1. An aposteriori eigenvalue error estimate is given via the residual

$$\frac{\|(A - \tilde{\lambda}_i B)\,\tilde{v}_i\|}{\|B\tilde{v}_i\|}$$

where $\tilde{\lambda}_i$, $\tilde{v}_i$ is the ith computed eigenvalue and eigenvector, respectively. CPU times are given but again, as in the least-squares problem, programming optimization was not used. An analysis of the convergence properties of the algorithm will be the subject of a future paper.

Example 1 (Gregory and Karney [5]).

$$A = \begin{bmatrix} 6 & 4 & 4 & 1 \\ 4 & 6 & 1 & 4 \\ 4 & 1 & 6 & 4 \\ 1 & 4 & 4 & 6 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$x_0^* = [-3 \quad 1 \quad -1 \quad 7] \qquad \text{(initial vector)}$$

no. of iterations	λ_i (true)	λ_i (computed)	Residual
20	5	5	.71720*E*-06
15	5	5	.82127*E*-04
19	15	15	.12561*E*-05
2	−1	−1	.20654*E*-06

Computed Eigenvectors

v_3	v_4	v_1	v_2
.49999999	.49999997	−.69337521	.13867504
.50000001	−.50000005	.13867504	.69337523
.50000001	−.50000003	−.13867505	−.69337526
.49999999	.49999994	.69337527	−.13867505

Note that $\lambda = 5$ is of multiplicity 2
CPU time = 1.2 sec.

EXAMPLE 2 (Golub [4]).

$$A = \begin{bmatrix} 42 & 48 & 27 & 8 & 1 & 0 \\ 48 & 69 & 56 & 28 & 8 & 1 \\ 27 & 56 & 70 & 56 & 28 & 8 \\ 8 & 28 & 56 & 70 & 56 & 27 \\ 1 & 8 & 28 & 56 & 69 & 48 \\ 0 & 1 & 8 & 27 & 48 & 42 \end{bmatrix}$$

$$B = \begin{bmatrix} 10 & 6 & 1 & 0 & 0 & 0 \\ 6 & 11 & 6 & 1 & 0 & 0 \\ 1 & 6 & 11 & 6 & 1 & 0 \\ 0 & 1 & 6 & 11 & 6 & 1 \\ 0 & 0 & 1 & 6 & 11 & 6 \\ 0 & 0 & 0 & 1 & 6 & 10 \end{bmatrix}$$

$x_0^* = [1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1]$ (initial vector)

no. of iterations	λ_i (true)	λ_i (computed)	Residual
23	9.06122577	9.06122577	.89699*E*-06
3	3.01131503	3.01131503	.15162*E*-06
1	0.10462927	0.10462922	.14010*E*-05
58	6.16251361	6.16251361	.18911*E*-05
3	0.00107213	0.00107896	.64117*E*-02
4	0.89558585	0.89558584	.98944*E*-04

Computed Eigenvectors

v_1	v_2	v_3
.231920684	.521120890	.417906545
.417906397	.231920613	−.521120885
.521120946	−.417906506	.231920552
.521120946	−.417906506	.231920552
.417906397	.231920613	−.521120885
.231920684	.521120890	.417906545

v_4	v_5	v_6
.417907061	−.231854922	−.521229448
.521119272	.417872469	.232110462
.231921861	−.521155233	.417661590
−.231922630	.521171522	−.41766906
−.521119889	−.417893509	−.232117435
−.417907403	.231880094	.521226857

CPU time = 4.6 sec.

REFERENCES

1. E. K. BLUM, "Numerical Analysis and Computation," Addison–Wesley, Reading, MA, 1972.
2. E. K. BLUM, Stationary points of functionals in a pre-Hilbert space, *J. Comp. Sys. Sci.* **1** (1967), 86–90.
3. A. A. GOLDSTEIN, "Constructive Real Analysis," Harper and Row, New York, 1967.
4. G. GOLUB, personal communication.
5. R. T. GREGORY AND D. L. KARNEY, "A Collection of Matrices for Testing Computational Algorithms," Wiley-Interscience, New York, 1969.
6. L. V. KANTOROVICH AND G. P. AKILOV, "Functional Analysis in Normed Spaces," MacMillan, New York, 1964.
7. S. F. MCCORMICK, A general approach to one-step iterative methods with application to eigenvalue problems, *J. Comp. Sys. Sci.* **6** (1972), 354–372.
8. S. F. MCCORMICK AND G. H. RODRIGUE, Gradient Methods for least-squares solution of linear operator equations, in preparation.
9. M. Z. NASHED, Steepest descent for singular linear operator equations, *SIAM J. Numer. Anal.* **7** (1970), 479–492.
10. P. WOLFE, Convergence conditions for ascent methods, *SIAM Rev.* **11** (1969), 226–235.
11. G. H. RODRIGUE, A variational method for the numberical solution of algebraic problems, Ph.D. dissertation, University of Southern California, Los Angeles, 1971.
12. A. E. TAYLOR, "Introduction to Functional Analysis," Wiley, New York, 1967.
13. J. H. WILKINSON, "The Algebraic Eigenvalue Problem," Oxford University Press, London, 1965.

14. G. PETERS AND J. H. WILKINSON, $Ax = \lambda Bx$ and the Generalized Eigenvalue Problem, *SIAM J. Numer. Anal.* **7** (1970).
15. W. J. KAMMERER AND M. Z. NASHED, On the convergence of the conjugate gradient method for singular linear operator equations, *SIAM J. Numer. Anal.* **9** (1972).
16. C. B. MOLER AND G. W. STEWART, An algorithm for the generalized matrix eigenvalue problems, *SIAM J. Numer. Anal.* **10** (1973).
17. G. H. GOLUB, R. UNDERWOOD, AND J. H. WILKINSON, The Lanczos Algorithm for the Symmetric $Ax = \lambda Bx$ Problem, Stanford U. Report STAN-CS-72-270, Mar. 1972.